Nucleic Acids Research, 2020 **1**
doi: 10.1093/nar/gkaa464

Comprehensive database and evolutionary dynamics of U12-type introns

Devlin C. Moyer[1,*,†], Graham E. Larue[2,*,†], Courtney E. Hershberger[1], Scott W. Roy[3,*] and Richard A. Padgett[1,*]

[1]Department of Cardiovascular and Metabolic Sciences, Lerner Research Institute, Cleveland Clinic Lerner College of Medicine, Cleveland Clinic and Department of Molecular Medicine, Case Western Reserve University, Cleveland, OH 44106, USA, [2]Department of Molecular and Cell Biology, University of California, Merced, Merced, CA 95343, USA and [3]Department of Biology, San Francisco State University, San Francisco, CA 94132, USA

Received April 07, 2020; Revised May 19, 2020; Editorial Decision May 19, 2020; Accepted May 20, 2020

ABSTRACT

During nuclear maturation of most eukaryotic pre-messenger RNAs and long non-coding RNAs, introns are removed through the process of RNA splicing. Different classes of introns are excised by the U2-type or the U12-type spliceosomes, large complexes of small nuclear ribonucleoprotein particles and associated proteins. We created intronIC, a program for assigning intron class to all introns in a given genome, and used it on 24 eukaryotic genomes to create the Intron Annotation and Orthology Database (IAOD). We then used the data in the IAOD to revisit several hypotheses concerning the evolution of the two classes of spliceosomal introns, finding support for the class conversion model explaining the low abundance of U12-type introns in modern genomes.

INTRODUCTION

The process of RNA splicing is a necessary step in the maturation of nearly all eukaryotic pre-messenger RNAs and many long non-coding RNAs. During this process, introns are excised from primary RNA transcripts, and the flanking exonic sequences are joined together to form functional, mature messenger RNAs (1,2). In most organisms, introns can be excised through two distinct pathways: by the major (>99% of introns in most organisms) or minor (<1% in most organisms, with some organisms lacking minor class introns altogether) spliceosomes. Despite the existence of eukaryotic species lacking the minor spliceosome, many reconstructions have shown that all eukaryotes descended from ancestors that contained minor class introns in their genomes, all the way back to the last eukaryotic common ancestor (3,4). The minor class introns have consensus splice site and branch point sequences distinct from the major class introns (5,6). It was originally thought that the two classes of introns were distinguished by their terminal dinucleotides, with introns recognized by the major spliceosome beginning with GT and ending with AG, and introns recognized by the minor spliceosome beginning with AT and ending with AC. However, it was later shown that introns in both classes can have either sets of terminal dinucleotides and that longer sequence motifs recognized by the snRNA components unique to each spliceosome distinguish the two classes of introns, hence the designations of 'U2-type' for the major and 'U12-type' for the minor spliceosomes (7).

The large-scale and well-organized online databases of genomic data, like Ensembl (8), UCSC (9) and RefSeq (10), do not provide extensive annotation information of intronic sequence in particular. Many databases focusing primarily on intron annotation information were created in the early 2000s, but most are no longer accessible (11–16), and the ones that remain accessible have not been updated in many years (17,18). The Exon-Intron Database (EID) (14) was one of the most comprehensive and robust databases in this group, and served as a basis for many further investigations into the peculiarities of introns (19–21), including other, more niche intron annotation databases (15,22). EID was maintained for at least six years, as it was updated in 2006 (23), but it is no longer accessible. Some more recent databases have been created, like ERISdb (24), JuncDB (25) and MIDB (26), but they are relatively narrow in scope: ERISdb only annotates splice sites in a selection of plant genomes; JuncDB annotates splice sites in a wide variety of genomes, but does not have any other easily-accessible intron annotation information; MIDB only annotates U12-type introns in the human and mouse genomes. Of all of

*To whom correspondence should be addressed. Tel: +1 831 420 7720; Email: egrahamlarue@gmail.com
Correspondence may also be addressed to Devlin C. Moyer. Email: devmoy@gmail.com
Correspondence may also be addressed to Scott W. Roy. Email: scottwroy@gmail.com
Correspondence may also be addressed to Richard A. Padgett. Email: padgetr@ccf.org
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

© The Author(s) 2020. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/nar/advance-article-abstract/doi/10.1093/nar/gkaa464/5850313 by guest on 05 June 2020

the databases mentioned above, U12DB (18), ERISdb and MIDB are the only databases that annotate intron class. Since U12DB has very old annotation data and ERISdb and MIDB only annotate introns in a small number of genomes, there is presently no publically available source of current U12-type intron annotation for an evolutionarily diverse array of organisms.

Many features of eukaryotic introns have been examined for clues about their evolutionary history. Introns can be assigned to one of three phases based on their position relative to the codons of the flanking exonic sequence: phase 0 introns fall directly between two codons, phase 1 introns fall between the first and second nucleotides of a single codon, and phase 2 introns fall between the second and third nucleotides of a single codon. It has long been noted that introns are not evenly distributed between the three phases (27,28). In conjunction with sequence biases on the exonic sides of splice sites, the phase biases were frequently cited by both sides of the debate between the proponents of the 'exon theory of genes' (the idea that primordial genes arose through exon shuffling and introns originally came into existence to facilitate this) (29) and those who argued that spliceosomal introns are descended from group II introns that invaded the ancestral eukaryotic genome, preferentially inserting themselves into so-called 'proto-splice sites' (30–32). Shortly after the discovery of U12-type introns (5), it was noted that the distribution of U12-type introns in the human genome was nonrandom, further complicating the debate around models explaining the origins of introns by requiring them to explain the presence of two classes of introns, the large discrepancy in the numbers of introns in each class, and the nonrandom distribution of U12-type introns (33). Furthermore, the phase biases in U12-type introns were noted to be different from the previously-documented phase biases in U2-type introns (34,35).

In an effort to address some of the many open questions about intron evolution, we created the Intron Annotation and Orthology Database (IAOD), a database of intron information for all annotated introns in 24 genomes, including plant, fungal, mammalian, and insect genomes. It also uniquely annotates orthologous introns, and assigns intron class using the intronIC algorithm described herein (Materials and Methods). The website is publicly accessible at introndb.lerner.ccf.org.

MATERIALS AND METHODS

Classifying intron type with intronIC

To begin, intronIC identifies all intron sequences in an annotation file by interpolating between coding features (CDS or exon) within the longest isoform of each annotated gene. For each intron, sequences corresponding to the 5′ splice site (5′SS, from −3 to +9 relative to the first base of the intron) and branch point sequence (BPS) region (from −55 to −5 relative to the last base of the intron) are scored using a set of position weight matrices (PWMs) representing canonical sequence motifs for both U2-type and U12-type human introns. A small 'pseudo-count' frequency value of 0.001 is added to all matrix positions to avoid zero division errors while still providing a significant penalty for low-frequency bases. For all scored motifs, the binary logarithm of the U12/U2 score ratio (the log ratio) is calculated, resulting in negative scores for introns with U2-like motifs, and positive scores for introns with U12-like motifs. Because U2 introns are not known to contain an extended BPS motif, the PWM for the U2-type BPS is derived empirically using the best-scoring U12 BPS motifs from all introns in the final dataset whose 5′SS U12 scores are below the 95th percentile. To identify the most likely BPS for each intron, all 12-mer sequences within the BPS region are scored and the one with the highest U12 log ratio score is chosen. This initial scoring procedure follows the same general approach used by a variety of different groups for bioinformatic identification of U12-type introns (4,33–36).

As originally shown by Burge *et al*., the 5′SS and BPS scores together are sufficient to produce good binary clustering of introns into putative types, due to strong correspondence between the 5′SS and BPS scores in U12-type introns (4,33). While this general feature of the data has often been employed in the identification of U12-type introns, a variety of different techniques have been used to define the specific scoring criteria by which an intron is categorized as U2- or U12-type. Here, we have implemented a machine learning method which uses support vector machine (SVM) classifiers (37) to assign intron types, an approach which produces good results across a diverse set of species and provides an easy-to-interpret scoring metric.

Our classification method relies upon two pieces of data: PWMs describing sequence motifs for the different subtypes of U2-type (GT-AG/GC-AG) and U12-type (GT-AG/AT-AC) introns, and sets of high-confidence U2- and U12-type intron sequences with which to train the SVM classifier (Figure 1A). Due to the scarcity of bona fide, experimentally-verified U2- and U12-type introns, a certain amount of curation was required to compile type-specific classifier training and scoring data. For the U12-type set, introns from six previously-published studies (18,38–42) as well as highly-conserved introns from a number of multi-species ortholog alignments were scored using SpliceRack (34) PWMs, and those with 5′SS scores >0 (i.e. 5′SS motifs more similar to U12-type than U2-type) present in at least three different sources were kept for use as U12-type training data. Combining these introns with branch point data from (41), we identified likely U12-type BPS motifs which were then used to generate BPS PWMs, requiring an A at either position +9 or +10 (following (35)). For the U2-type set, we first collected intron sequences from the yeast *Schizosaccharomyces pombe*, a species which is believed to lack U12-type introns. These introns were then filtered using data from (43) to include only those with direct evidence of splicing, and scored against human SpliceRack PWMs to establish an upper bound for SpliceRack U12 PWM scores on high-confidence U2-type introns. Finally, using a set of introns conserved between human, zebrafish and horseshoe crab we identified human introns found in orthologous groups where every constituent intron had a 5′SS SpliceRack PWM score less than the *S. pombe* U2-type threshold. These human U2-type introns were combined with the U12-type set to build an updated collection of PWMs, and to define positive (U12-type) and negative (U2-type) training sequences for the SVM (Figure 1C). In order to establish U2-type BPS PWMs specific to each unique set of input

Downloaded from https://academic.oup.com/nar/advance-article-abstract/doi/10.1093/nar/gkaa464/5850313 by guest on 05 June 2020

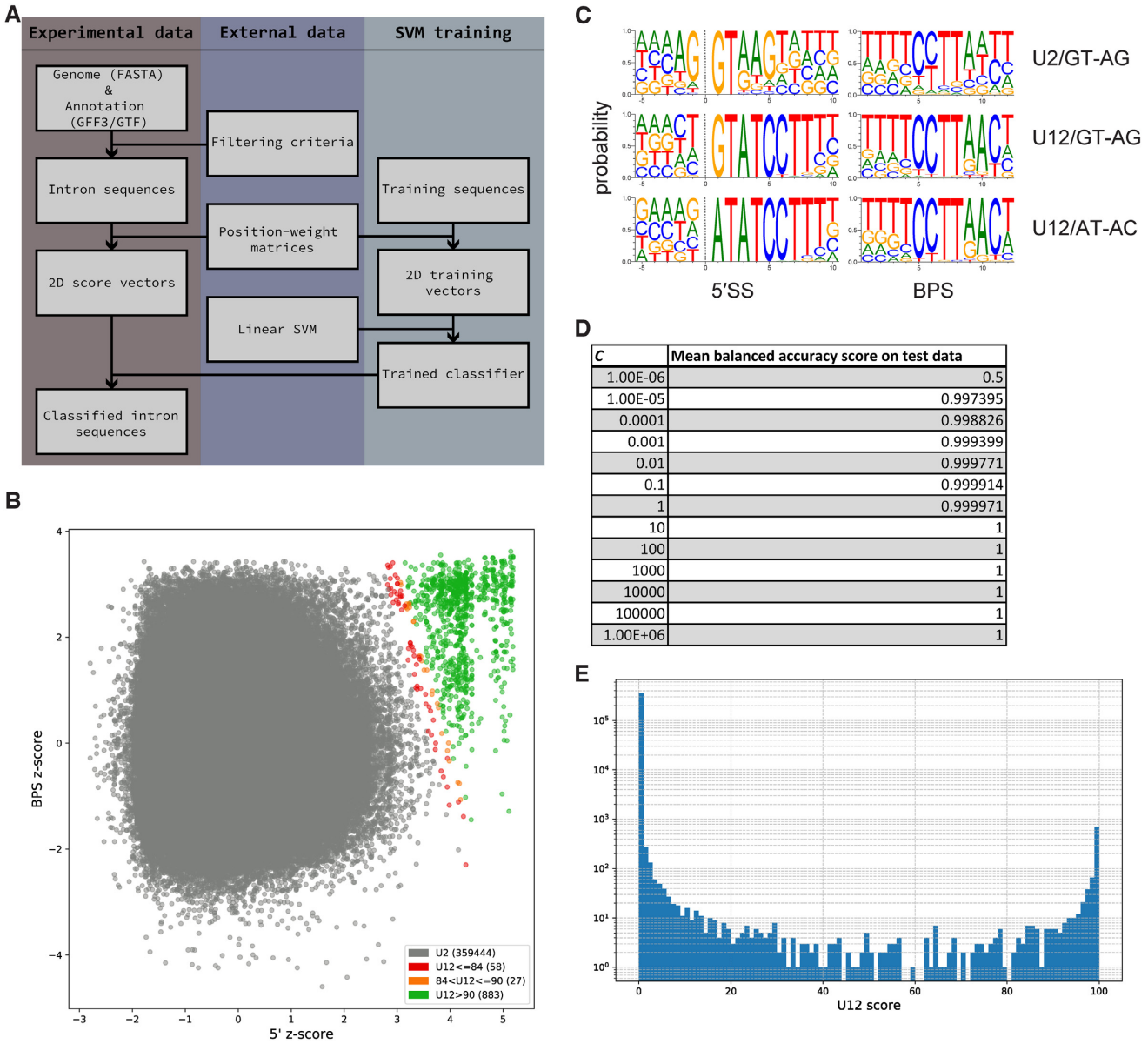

Figure 1. (**A**) Overview of the major steps of the intronIC algorithm. (**B**) Scatter plot of all classified introns in the human genome; gray: U2-type introns, red: U12-type introns with probability scores ≤84%; yellow: U12-type introns with probability scores from 84–90%, green: U12-type introns with probability scores >90%, our chosen scoring threshold. (**C**) Sequence logos of the 5′SS and BPS PWMs for GT-AG/AT-AC U2- and U12-type introns. (**D**) Balanced accuracy performance of the classifier with different values of hyperparameter *C* on test sets during the first round of the cross-validation process. (**E**) Histogram (with logarithmic scale y-axis) of probability scores for the human data shown in part B.

introns (e.g. each different species), U12-type PWMs are first used to find the highest-scoring BPS motifs for all introns whose 5′SS U12-type scores are lower than the 95th percentile (i.e. introns unlikely to be U12-type). These sequences are then used to create U2-type BPS PWMs, making the overall BPS scoring more conservative by defining the U2-type BPS PWMs using the most U12-like BPS motifs found in the empirical data (similar to the approach described in (4)).

Because clear discrimination between U12-type and U2-type introns can be achieved by considering only two scoring dimensions (4,33,35), we use a relatively simple SVM classifier with a linear kernel as implemented in the scikit-learn Python library (44). The SVM is trained on a set of two-dimensional vectors, corresponding to the 5′SS and BPS scores of the introns in the training data, which are labeled by intron type. For linear classifiers there is only a single free hyperparameter to be adjusted, *C*, which is (roughly) the degree to which misclassification of data in the training set is penalized during the creation of an optimized (i.e. wide) margin separating the positive and negative classes. To our knowledge there is no single, standard approach for establishing the best value of *C*; we chose to optimize *C* using an iterative cross-validation method which starts with a wide range of logarithmically-distributed values and narrows that range based upon the best-performing

Downloaded from https://academic.oup.com/nar/advance-article-abstract/doi/10.1093/nar/gkaa464/5850313 by guest on 05 June 2020

(highest balanced accuracy score) value of C in each validation round. After several such iterations, the mean of the resulting range is taken as the final value of C to be used to train the classifier. Balanced accuracy is used as a performance metric due to the highly imbalanced nature of the training data, where the negative class (U2-type) greatly outnumbers the positive class (U12-type). Because the human training data is very well-separated, when applied to intron sequences in the human genome values of $C \geq 10$ perform equally well during cross-validation (Figure 1D). Given the broad range of good parameter values, taking the average of all best-performing values results in a more conservative margin (larger C) than taking the default 'best' parameter value via the scikit-learn API, which simply returns the first rank-1 parameter value found. For the human genome, this approach results in a classifier which performs perfectly on the training sets, with both F1 and precision-recall AUC scores of 1.0 on held-out training data (examples of final scores for human introns in Figure 1B,E).

For the purpose of populating the IAOD, intronIC was slightly modified to produce a single output file containing all of the annotation information recorded in the IAOD for each intron—the default version is available for download from (https://github.com/glarue/intronIC) and the modified version used for this application is available at (https://github.com/Devlin-Moyer/IAOD).

The annotation and sequence files provided as input to intronIC were downloaded from release 92 of Ensembl (with the exception of the FUGU5 assembly of the *Takifugu rubripes* genome, which was downloaded from release 94), release 39 of Ensembl Metazoa, or release 40 of Ensembl Plants (8). Data was obtained for every genome annotated by U12DB with the addition of *Zea mays*, *Oryza sativa*, *Glycine max* and *Schizosaccharomyces pombe* to increase the evolutionary diversity of the represented genomes.

Annotating introns in non-coding transcripts/regions

To annotate introns, intronIC can use either exon or CDS entries in a GFF3 or GTF file. When using exon entries to define introns, intron phase is undefined. In order to get complete annotation of both introns within open reading frames and within untranslated regions or non-coding transcripts, intronIC was run twice on each genome analyzed, once producing exon-defined introns and once producing CDS-defined introns. A custom Python script then compared both lists of introns to produce a single list where the CDS-defined intron annotation information was used if the intron was in a coding region and the exon-defined information was used otherwise.

Finding gene symbols

The output of intronIC includes the Ensembl gene ID but not the gene symbol for all introns in a genome using an annotation file from Ensembl. Ensembl maintains vast databases of genomic data which are accessible with BioMart (45). BiomaRt (46) is an R package for interacting with these databases. A custom R script submitted a list of all of the Ensembl gene IDs in each genome in the database to biomaRt and obtained gene symbols for all of those genes.

Assigning orthologous introns

Coding sequences for every annotated transcript in each of the 24 genomes were extracted and translated into their corresponding protein sequences. These sequences were aligned with DIAMOND (47) to identify sets of best reciprocal hits—considered orthologs going forward—between every pairwise combination of species, using an E-value cutoff of 10^{-10} and –min-orfs set to 1. Every pair of orthologous transcripts was then globally aligned at the protein level using Clustal W (v2.1; ref. (48)), and all introns in regions of good local alignment between pairs (≥40% matching amino acid sequence ±10 residues around each intron) were extracted using custom Python scripts (following the approach of ref. (49)). Lastly, conservative clustering of the pairwise orthologous intron sets was performed through identification of all complete subgraphs where every member is an ortholog of every other member (i.e. maximal clique listing) to produce the final intron groups (e.g. A–B, A–C, B–C, B–D → A–B–C, B–D).

Database creation

A custom Python script created a PostgreSQL database using the output of intronIC, the lists of gene symbols from BioMart, and the list of orthologous groups of introns. All of the orthologous groups were inserted in a table with two columns: a unique numeric ID for each group and a list of all intron labels belonging to that group. One table for each genome contains, for each intron: the abbreviated sequence (see above), taxonomic and common names of the organism, name of the genome assembly, intronIC score, intron class (determined from the intronIC score), intronIC label, chromosome, start coordinate, stop coordinate, length, strand, rank in transcript, phase, terminal dinucleotides, upstream exonic sequence (50 nt), 3′ terminus with the branch point region enclosed with brackets (40 nt), downstream exonic sequence (50 nt), full intron sequence, Ensembl gene ID, Ensembl transcript ID, and gene symbol. Another table with identical fields contains all U12-type introns from all genomes.

Website design

The website was constructed using Django 2.0, an open-source Python web development framework, and Bootstrap 4.0.0, an open-source framework for front-end web development. The search engines use the Django ORM to interact with the PostgreSQL database.

There are four search engines on the website: the main, advanced, U12, and orthologous searches. The main and advanced search interfaces have input fields corresponding to individual columns in the database, so the text input in each field can easily be matched with the appropriate column using the Django ORM. The U12 search engine uses PostgreSQL search vectors to allow users to make full text queries against the database. I.e. users can input a string containing one term corresponding to as many fields as they like and get a result. However, if the search query contains, e.g., the names of two different species or genes, no results will be returned, since no single record (intron) in the database corresponds to multiple species or genes. This

Downloaded from https://academic.oup.com/nar/advance-article-abstract/doi/10.1093/nar/gkaa464/5850313 by guest on 05 June 2020

limits the number of possible queries, but allows for a simple user interface for simple queries concerning U12-type introns. Since the main and advanced search engines require users to specify which field of the database each term of their query corresponds to, it does not need to use full text search vectors, and can consequently accept multiple search terms for each field. The homolog search engine also makes use of PostgreSQL search vectors to find the row of the homolog table containing the intron ID input by the user.

Assessing randomness of distribution of U12-type introns

If U12-type introns were randomly inserted a genome, we would expect the distribution of U12-type introns per gene to be binomial with parameters n = number of genes with at least one U12-type intron and $p = 1 - (1 - x)^{m-1}$, where x is the proportion of U12-type introns in the genome and m is the average number of introns in the genome. Data from the IAOD was used to obtain n, x, and m for each genome in the database that contained at least one U12-type intron. The dbinom function in R was used to compute the probability of observing the observed number of genes with multiple U12-type introns in each genome. Supplementary Table S1 lists the parameters passed to the dbinom function.

To ensure that the observed clustering of U12-type introns in the same genes was not an artefact of U12-type introns with alternative splice sites being recorded as distinct U12-type introns, all intron coordinates listed by intronIC were used to create a graph where each node corresponded to a position within each genome (e.g. GRCh38 + chr1 + 492045 corresponds to base pair 492045 on chromosome 1 in assembly GRCh38 of the human genome) and two nodes are joined with an edge if they appear in the same row of the list of intron coordinates. In this graph, alternatively spliced introns are evident as clusters of >2 nodes, so each cluster represents a single intron, regardless of how many alternative splice sites it possesses. A single edge from each cluster was selected and the corresponding coordinates were matched to the original intronIC output to get accurate counts of the total number of unique introns in each class in all genomes annotated in the IAOD.

RESULTS AND DISCUSSION

Identification of U12-type introns across 24 eukaryotic species

We developed the method intronIC (see Materials and Methods), implemented in Python, and used it to perform genome-wide identification of U2- and U12-type spliceosomal introns in 24 eukaryotic species including 14 vertebrate animals, 5 invertebrate animals, 4 plants and two yeasts. For each species, type-specific position-weight matrices (PWMs) for the 5′ splice site and branch point sequences were used to create score vectors for every intron in each genome. These score vectors were then compared against corresponding vectors in high-confidence training sets from *Homo sapiens* using a machine-learning (SVM) classifier to assign each intron a probability of being U12-type (see Figure 1 for an overview of the major steps in the algorithm and examples of classifier performance on human data). Once trained on the conserved intron data, the classifier assigned every intron in the experimental set a probability of being U12-type. Introns with at least a 90% probability of being U12-type were classified as U12-type, which produces classifications in good agreement with previously-reported findings for well-studied species. For example, running our method on U12-type intron sequences from the U12DB (18) results in equivalent classifications for 96% (381/398) of the U12DB introns in chicken, 97% (535/554) in mouse, 94% (15/16) in *Drosophila melanogaster* and 95% (656/691) in human. In *Arabidopsis thaliana*, our method matches the calls in the U12DB 94% of the time (223/238), with similar results (269/292, 92%) for U12-type introns from the plant-specific database ERISdb (24). Furthermore, in each test species listed above intronIC identifies additional putative U12-type introns not present in existing databases, likely due to a combination of newer annotation data and our method's sensitivity. In *Caenorhabditis elegans*, a well-annotated species believed to have lost all of its U12-type introns, when run on all introns (not just those from the longest isoform per gene) our method categorized only 1/116241 introns as U12-type, suggesting a false-positive rate of less than 0.001%. A total of 8967 U12-type introns were identified using this technique. Groups of analyzed introns in conserved regions of homologous genes were also annotated; collectively, these data constitute the Intron Annotation and Orthology Database (IAOD).

Figure 2 compares the number of U12-type introns annotated in each species in the IAOD with the numbers of U12-type introns annotated by previous databases annotating U12-type introns: U12DB (18), SpliceRack (35) and ERISdb (24). The IAOD often annotates many more introns than U12DB, likely due to the different approaches to annotating intron class and the quality of the genome assemblies used. In U12DB, U12-type introns were annotated by mapping a set of reference introns from *Homo sapiens*, *D. melanogaster*, *A. thaliana* and *Ciona intestinalis* to the whole genomes of every other organism in the database (18), while introns in the IAOD were annotated directly from every genome in the database using the intron-classifying program intronIC (see Materials and Methods for details). U12DB primarily annotates U12-type introns in all represented species that are orthologous to the reference U12-type introns (18), while the IAOD annotates U12-type introns in all genomes independently, using the species-specific annotations for each genome. Furthermore, the genome assemblies and annotations used to identify introns in the present study are all several versions newer than those used in U12DB, so part of the discrepancy in the number of U12-type introns annotated is likely due to an increase in the number of annotated genes and splice sites since the creation of U12DB. While intronIC itself does not provide homology information about the annotated introns, the IAOD also annotates intron orthologs: of the 3 645 636 total introns in the IAOD, 54% (1 989 840) have at least one other intron annotated as being in a conserved region of a homologous gene in another genome in the IAOD.

As shown in Figure 2, there are substantially fewer U12-type introns in the analyzed invertebrate animals than in the vertebrates, and none in either species of yeast analyzed, consistent with earlier findings (33,36). The numbers

Downloaded from https://academic.oup.com/nar/advance-article-abstract/doi/10.1093/nar/gkaa464/5850313 by guest on 05 June 2020

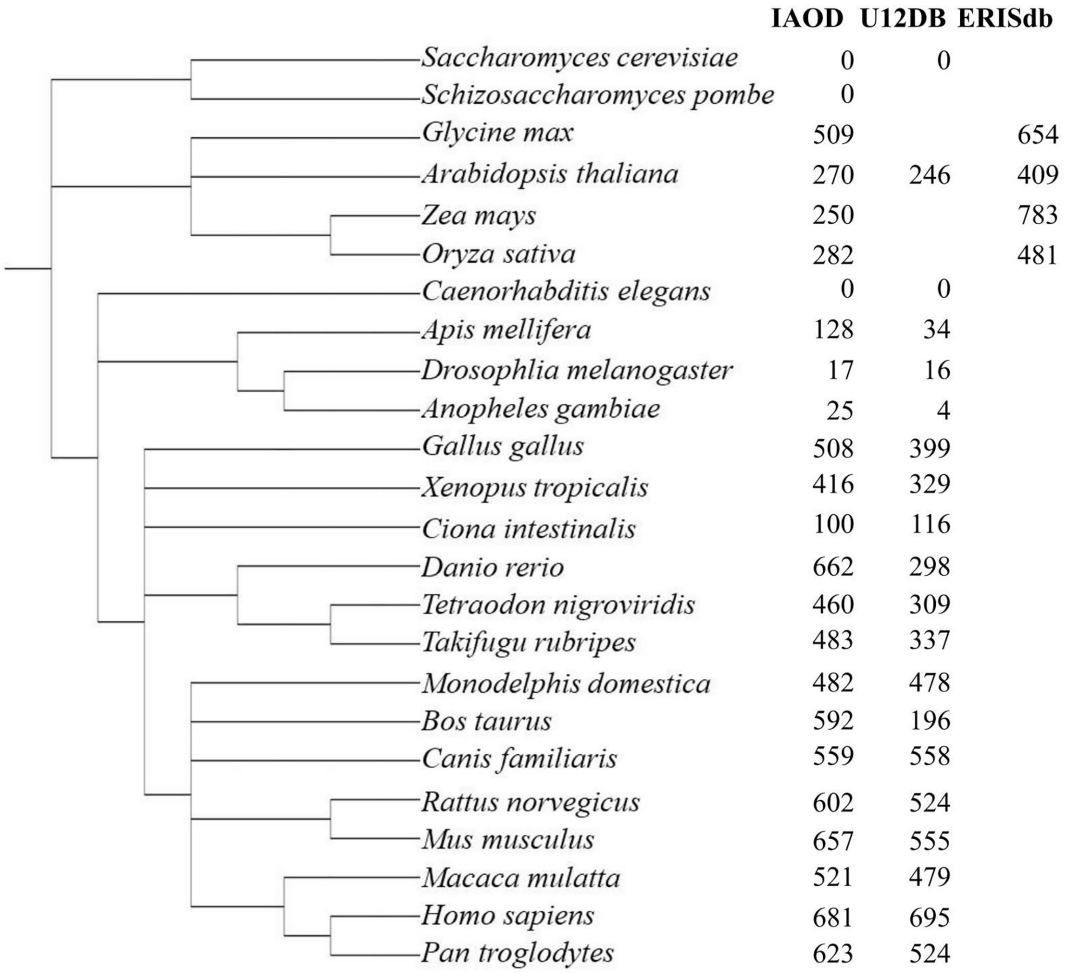

Figure 2. Phylogenetic distribution of U12-type introns in all species annotated by the Intron Annotation and Orthology Database (IAOD), U12DB (18), SpliceRack (35) and ERISdb (24). Blank entries in the table represent organisms not represented in the respective database. Counts of U12-type introns in the IAOD only represent introns flanked by coding exons. The NCBI Taxonomy Browser (70) and Integrative Tree of Life (71) were used to create the phylogenetic tree.

of introns determined by intronIC to be U12-type in a few species deserve special attention. The numbers of U12-type introns in *A. thaliana*, *O. sativa* and *Z. mays* are noteworthy because there are substantially fewer U12-type splice sites annotated in the IAOD than in ERISdb, but inspection of the U12-type splice sites annotated in ERISdb reveals many duplicate sequences. These duplicates arise from the fact that ERISdb counts each set of U12-type splice sites from every transcript of every gene as a distinct set of U12-type splice sites. In the case of *A. thaliana*, of the 414 U12-type splice sites annotated in ERISdb, there are only 292 unique sequences, which is much closer to the 269 annotated in the IAOD.

Phase bias of U12 introns is consistent with the conversion hypothesis

The phase biases observed in the IAOD (Figure 3) agree with the results of previous studies and extend them to many more organisms: an excess of phase 0 introns among U2-type introns (27,28,33,34,50), and a bias against phase 0 introns among U12-type introns (33,35) are seen in all studied lineages, and the presence of these biases in both plant and animal genomes suggest a deep evolutionary source. Multiple explanations for the overrepresentation of phase 0 U2-type introns have been proposed, including exon shuffling (51), insertion of introns into proto-splice sites (30,50), and preferential loss of phase 1 and 2 introns (6). These models do not consider or explain the underrepresentation of phase 0 U12-type introns.

To account for the phase biases present in U12-type introns, we propose that the observed phase biases in both classes of introns can be explained by an extension of the class-conversion hypothesis proposed by Burge *et al.* (33). This hypothesis arose from the observation that U12-type introns in human genes were often found to have U2-type introns at orthologous positions in *C. elegans* genes. Dietrich *et al.* (7) showed that U12-type introns could be converted to U2-type introns with as few as two point mutations. These results also suggest that class conversion is likely to only proceed from U12-type to U2-type, a hypothesis for which we find support in the distinctly U12-like phase distribution among U2-type introns with >1 U12-type ortholog (i.e. putative U12-type-to-U2-type conver-

Downloaded from https://academic.oup.com/nar/advance-article-abstract/doi/10.1093/nar/gkaa464/5850313 by guest on 05 June 2020

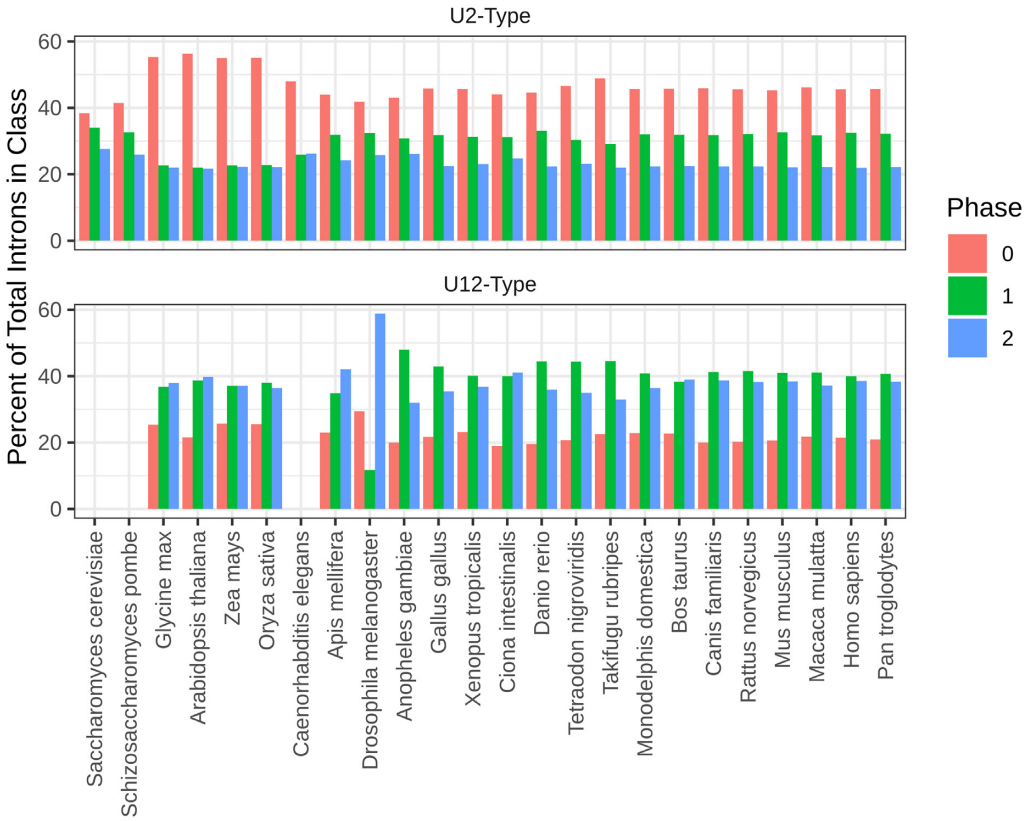

Figure 3. Phase distribution of introns within each class in all genomes annotated in the IAOD. Organisms are grouped by phylogeny. The bias against phase 0 U12-type introns is statistically significant in all organisms but *G. max*, *O. sativa*, *X. tropicalis* and *Z. mays* (chi-squared; $P < 0.10$). The bias toward phase 0 U2-type introns is statistically significant in all organisms but *A. mellifera*, *D. melanogaster*, *S. cerevisiae* and *S. pombe* (chi-squared; $P < 0.10$).

sions) (Supplementary Figure S1). In light of this, one possible explanation for the current data is that, at an early stage in eukaryotic evolution, there were many more U12-type introns than are currently observed in any characterized genome, and the phase bias arose as phase 0 U12-type introns were preferentially converted into U2-type introns, producing both an overrepresentation of phase 0 U2-type introns and an underrepresentation of phase 0 U12-type introns. This selectivity for phase 0 introns in the class conversion process rests on the function of the –1 nucleotide relative to the 5′ splice site in both spliceosomes.

As shown in Figure 4, there is a large excess of G at the –1 position of U2-type 5′ splice sites, across all three phases, in agreement with earlier investigations (51,52). Figure 4 also shows that there is an excess of -1U in U12-type introns in all three phases. There also appears to be a bias against –1 A and G in phase 1 and phase 2 U12-type introns, but not in phase 0 U12-type introns. Interestingly, when introns are grouped by terminal dinucleotides, these biases are only found in U12-type introns with GT-AG terminal dinucleotides and not in U12-type AT-AC introns (Figure 5). The preference for –1G at U2-type 5′ splice sites appears to be due to the fact that the –1 nucleotide pairs with a C on the U1 snRNA (53,54). The preference for –1U at U12-type 5′ splice sites is more mysterious, as no snRNAs are known to bind to this position. It was previously shown that the U11/U12-48K protein interacts with the +1, +2 and +3 nucleotides at the U12-type 5′ splice site in a sequence-specific fashion, but the specificity of the interaction with the -1 position was not studied (55). As noted above, the bias against –1G in U12-type introns with GT-AG terminal dinucleotides could be a consequence of the gradual conversion of many U12-type introns into U2-type introns. The lack of consistent -1 nucleotide biases in AT-AC introns of either class may be due to the fact that AT-AC introns are poorly recognized by the U2-type spliceosome (54) and were thus largely unaffected by the class conversion process.

We propose that this preference for conversion of phase 0 U12-type introns is due to the fact that introns with a G at the -1 position relative to the 5′ splice site bind more strongly to the U1 snRNA (54,55), and the -1 nucleotide of phase 0 introns is the final wobble position of the corresponding codon and can be a G in 13 of 20 codon families. Thus, the -1 nucleotides of phase 0 U12-type introns were more free to mutate to G and increase the affinity of the U2-type spliceosome for their 5′ splice sites, gradually accumulating mutations in the other sequences required for recognition by the U12-type spliceosome (7). Table 1 contains some examples of orthologous introns of different classes that demonstrate the class conversion process. This unidirectional conversion process also provides an explanation for the low abundance of U12-type introns in modern eukaryotic genomes.

Downloaded from https://academic.oup.com/nar/advance-article-abstract/doi/10.1093/nar/gkaa464/5850313 by guest on 05 June 2020

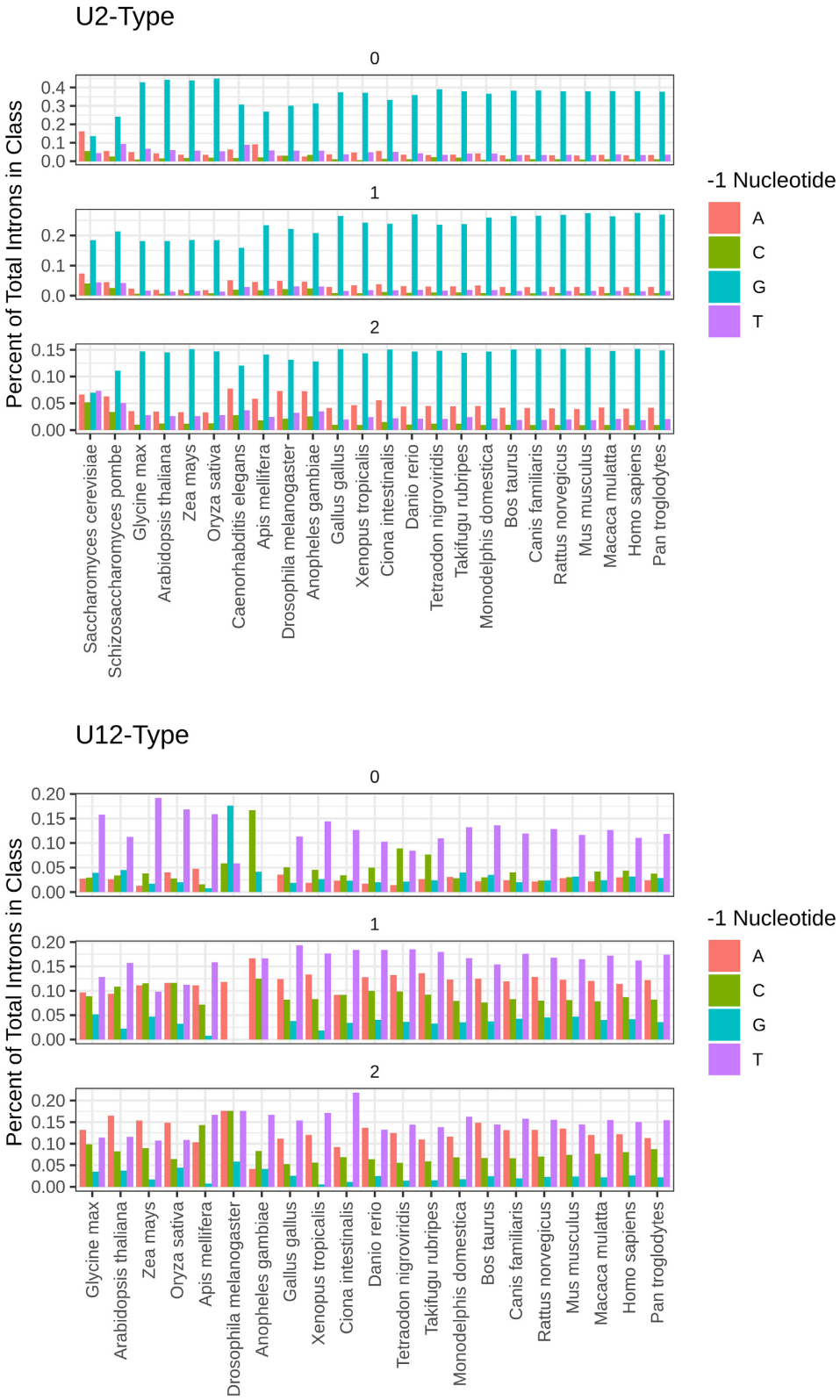

Figure 4. Percentages of introns with the specified nucleotide immediately upstream of the 5′ splice site in each phase of both classes of introns. Organisms are grouped by phylogeny.

Downloaded from https://academic.oup.com/nar/advance-article-abstract/doi/10.1093/nar/gkaa464/5850313 by guest on 05 June 2020

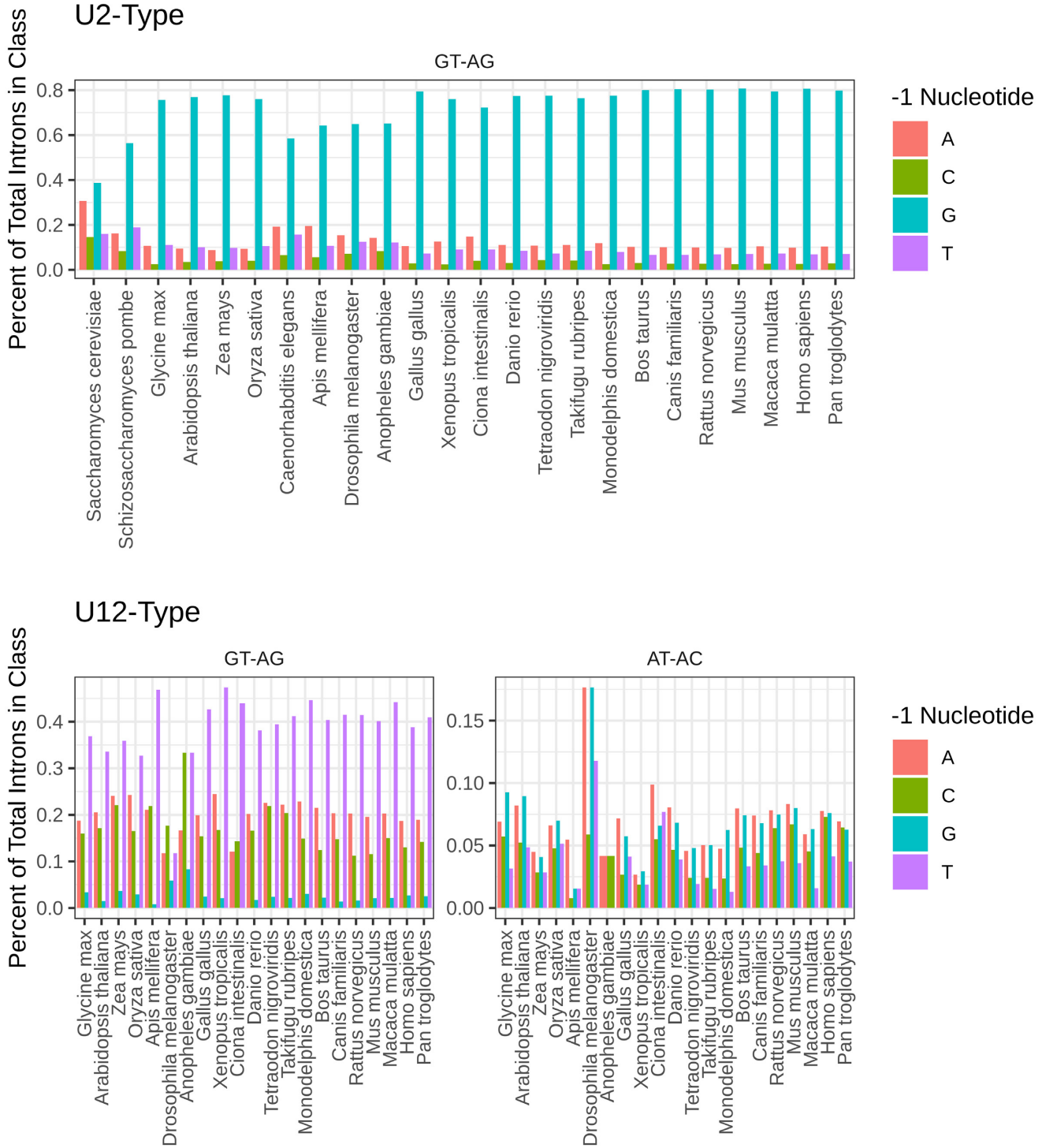

Figure 5. Nucleotide biases at the -1 position relative to the 5′ splice site for all organisms (excluding those lacking U12-type introns) annotated in the IAOD, grouped by terminal dinucleotides and intron class.

Downloaded from https://academic.oup.com/nar/advance-article-abstract/doi/10.1093/nar/gkaa464/5850313 by guest on 05 June 2020

Table 1. Groups of orthologous introns of different classes. Introns with nothing in the gene column came from transcripts with no gene name annotated by Ensembl. Vertical bars in the sequence column denote splice sites, and the middle sequence flanked by ellipses is the putative branch point region as annotated by intronIC

Organism	Gene	Intron Class	Phase	Sequence
Zea mays		U12-type	0	GCAAAG|GTATCCTTTT...TCCTCCTAAACT...TGCAG|TCCTCC
Oryza sativa		U2-type	0	GCCAAG|GTAATTTATA...TTAATGTTTAAT...TGCAG|TTAATG
Zea mays		U2-type	0	AAGCGG|GTATGTCTAG...TTGATCTCACCT...ATCAG|TTGATC
Glycine max		U12-type	0	AAGCGT|GTATCCTTCA...TTGTCCTTGACC...GAAAG|TTGTCC
Zea mays		U2-type	1	TCAACA|GTACGCAACA...TCCTTCTTAATT...TGTAG|TCCTTC
Oryza sativa		U12-type	1	TCAACA|GTATCCATCA...TTTTTCTTAACT...TGTAG|TTTTTC
Arabidopsis thaliana		U2-type	1	TCAACA|GTAAATTTTC...TTTCTCTTGACC...TGCAG|TTTCTC
Canis familiaris	*SMYD2*	U12-type	1	ACAAAT|ATATCCTTTA...CTTTCCTTGACA...AGCAC|CTTTCC
Homo sapiens	*SMYD2*	U12-type	1	ATAAAT|ATATCCTTTA...CTTTCCTTGACT...AGCAC|CTTTCC
Mus musculus	*Smyd2*	U12-type	1	ACAAAT|ATAACCTTTC...GTTTCCTTGACG...AGCAC|GTTTCC
Macaca mulatta	*SMYD2*	U2-type	1	ACAACT|GCCCTGATGG...GTTTCCTTGACT...CACAG|GTTTCC
Pan troglodytes	*SMYD2*	U12-type	1	ATAAAT|ATATCCTTTA...GTTTCCTTGACT...AGCAC|GTTTCC
Rattus norvegicus	*Smyd2*	U12-type	1	ACAAAT|ATAACCTTTC...GTTTCCTTGACG...AGCAC|GTTTCC
Anopheles gambiae		U2-type	2	TAATCC|GTATGTAACC...TGTTTCTCCTTT...TGTAG|TGTTTC
Monodelphis domestica	*RNF121*	U12-type	2	TAACCC|GTATCCTTTT...TTTTCTTTAACC...TGAAG|TTTTCT
Rattus norvegicus	*Rnf121*	U12-type	2	CAATCC|GTATCCTTTG...TGATCCTTAACA...GACAG|TGATCC
Homo sapiens	*UFD1*	U12-type	2	AGCCGT|GTATCTTTTT...GTTGCCTTGACA...TGCAG|GTTGCC
Pan troglodytes	*UFD1*	U12-type	2	AGCCGT|GTATCTTTTT...GTTGCCTTGACA...TGCAG|GTTGCC
Tetraodon nigroviridis	*ufd11*	U2-type	2	AGCAGT|GTAAGAACGA...GAATTGTTTTCT...TGCAG|GAATTG

U12 introns are non-randomly distributed across genes

Multiple previous surveys of U12-type introns have revealed that the distribution of U12-type introns in the human genome is non-random, i.e. there is a statistically significant tendency for U12-type introns to cluster together in the same genes (33–35). Repeating this analysis for all genomes in the IAOD (except *S. cerevisiae*, *S. pombe* and *C. elegans*, as they lack U12-type introns) replicated their findings in all 21 genomes ($P < 0.05$ for all genomes; see Methods and Supplementary Table S1). Many explanations for this nonrandom distribution have been proposed, including the fission–fusion model of intron evolution (33); a difference in the speed of splicing of U12-type and U2-type introns (35,39); and the idea that the U12-type introns arose during an invasion of group II introns after U2-type introns had already seeded the ancestral eukaryotic genome, meaning the new U12-type introns could only be inserted in certain locations (56). The fission–fusion model posits that two separate lineages of the proto-eukaryote evolved distinct spliceosomes and then fused their genomes such that all genes originally contained either only U2-type introns or only U12-type introns. Thus, modern U2-type introns in genes also containing U12-type introns were originally U12-type introns that were subjected to the class conversion process discussed above (33). An alternative argument for the low abundance of U12-type introns is that they are excised more slowly than U2-type introns, so genes that contain U12-type introns contain them because those genes need to be expressed slowly for some reason (35,39). However, it has since been shown that the rate of excision of U12-type introns is not sufficiently different from the rate of excision of U2-type introns to produce a meaningful impact on the expression of transcripts containing U12-type introns (57). Furthermore, recent evidence suggests that the rates of both types of splicing are sufficiently fast that most introns will be excised cotranscriptionally (40).

Low conservation of U12-type intron positions between animals and plants

Basu *et al.* (58) argued that the number of U12-type introns present in the ancestral eukaryotic genome was unlikely to be substantially larger than the largest number of U12-type introns observed in any modern genome, thus suggesting that the process of class conversion is a minor evolutionary force. However, the basis of their argument is the finding that the positions of U12-type introns are more highly conserved than the positions of U2-type introns between humans and *Arabidopsis thaliana*, a result that the present data do not support: we find that out of the 93 U12-type introns in the human genome in regions of good alignment to *A. thaliana*, only 8 (9%) are in conserved positions, while out of the 9527 U2-type introns in such regions of the human genome, 2098 (22%) are in conserved positions in *A. thaliana*. Thus, our comparative analysis is consistent with U12-type intron enrichment in the ancestral eukaryotic genome relative to the most U12-intron-rich extant lineages.

Splicing boundaries of U12 introns

The majority of introns annotated in the IAOD in both classes begin with GT and end with AG (Table 2), in agreement with previous studies (17,33,35). A substantial minority of U2-type introns, but almost no U12-type introns, were found to have GC–AG as their terminal dinucleotides in many of the analyzed genomes, reflecting their previously documented role in alternative 5′ splice site selection

Downloaded from https://academic.oup.com/nar/advance-article-abstract/doi/10.1093/nar/gkaa464/5850313 by guest on 05 June 2020

Table 2. Percentages of introns with various terminal dinucleotides in each class in all annotated genomes. Organisms are sorted in the phylogenetic order shown in Figure 2

Intron class	U2-type				U12-type			
Terminal dinucleotides	GT-AG	GC-AG	AT-AC	Other	GT-AG	GC-AG	AT-AC	Other
Saccharomyces cerevisiae	912	2.7	0	5.7	0	0	0	0
Schizosaccharomyces pombe	100	0.12	0	0.01	0	0	0	0
Glycine max	98	1.6	0	0	76	0.01	23	0
Arabidopsis thaliana	99	1.0	0.01	0.08	73	0	26	1.1
Zea mays	99	0.50	0.05	0.41	86	0	13	1.2
Oryza sativa	97	0.30	0.01	2.5	74	0	22	3.8
Caenorhabditis elegans	99	0.64	0	0.18	0	0	0	0
Apis mellifera	99	0.65	0.01	0.03	91	0.71	8.6	0
Drosophila melanogaster	99	0.75	0.01	0.07	47	5.3	47	0
Anopheles gambiae	100	0.26	0	0.09	88	4.2	8.3	0
Ciona intestinalis	91	0.70	0.04	8.7	65	0	23	12
Gallus gallus	97	2.2	0.01	0.56	77	0.97	18	3.4
Xenopus tropicalis	85	0.90	0.07	14	78	0.22	8	14
Danio rerio	97	1.2	0.02	1.4	75	0	22	2.7
Tetraodon nigroviridis	86	1.4	0.03	13.0	77	0.80	12	11
Takifugu rubripes	91	5.7	0	3.4	79	4.3	12	5.0
Monodelphis domestica	98	0.50	0.01	1.1	82	0.20	14	4.1
Bos taurus	94	0.89	0.03	5.1	69	0.81	21	10
Canis familiaris	95	0.86	0.02	3.7	69	0.35	20	11
Rattus norvegicus	97	0.78	0.02	2.1	69	0.49	23	6.5
Mus musculus	99	0.78	0.01	0.16	69	0.47	25	5.4
Macaca mulatta	97	3.4	0.01	0.02	78	2.8	17	2
Pan troglodytes	97	2.8	0.01	0.14	72	1.7	21	4.7
Homo sapiens	99	0.77	0.01	0.15	68	0.61	26	5.2

in U2-type splicing in many organisms (6,35,59–61). Several previous studies have found numerous introns with other non-canonical terminal dinucleotides in multiple genomes, sometimes with functional roles in regulation of alternative splicing (17,35,59,62), but intronIC has annotated many thousands of U2-type introns with non-canonical terminal dinucleotides in certain organisms, such as *Gallus gallus* and *Tetraodon nigroviridis* (Table 2). Inspection of these introns reveals that the vast majority of these splice sites are only a few nucleotides away from a conventional U2-type splice site with canonical terminal dinucleotides; these splice sites with non-canonical dinucleotides were likely annotated on the basis of conserved exon boundaries, without regard for the precise placement of the splice sites. The proportion of U12-type introns with non-canonical terminal dinucleotides (Table 2) largely agrees with previous investigations (35,36,63).

Distribution of intron lengths of U12 and U2 introns

Supplementary Figure S2 shows the distributions of intron lengths in six of the genomes annotated in the IAOD, representing each general type of length distribution observed in the IAOD. In accordance with previous studies, when plotted on a log scale, there are two distinct peaks in the distribution of intron lengths in U2-type introns in humans and chicken while the distribution of U12-type intron lengths has only one peak (34,64) (these peaks are not apparent when length is plotted on a linear scale). A previous study considered the distribution of intron lengths amongst several eukaryotic genomes collectively (65), producing a distribution similar to those observed in the human and chicken genome in Supplementary Figure S2. However, Supplementary Figure S2 demonstrates great diversity in the distributions of intron lengths amongst eukaryotes; zebrafish have two distinct peaks of comparable size of intron lengths in both classes, while corn, honeybee and fugu have large peaks of shorter introns and very small peaks of longer introns in both classes. The significance of these variations is unclear; differing distributions of intron lengths in the two classes of introns have previously been used to argue that U12-type introns are recognized through intron definition, while U2-type introns are recognized by exon definition (66). However, Supplementary Figures S3 and S4 show that the mean intron lengths in both classes of intron in all 24 genomes annotated in the IAOD correlate strongly with genome size (Pearson's r: 0.87 for U12-type introns and 0.93 for U2-type introns), consistent with previous findings (64,65,67). This correlation suggests that mean intron lengths in both classes are generally a function of genome size and not a reflection of intron definition imposing a restriction on the size of U12-type introns.

Interestingly, Supplementary Figures S3 and S4 show that the relationship between mean intron length and genome size differs between vertebrates, insects, and plants. In insects, the total genome size remains very small and the mean intron length does not appear to correlate with total genome size. This may be related to the greater prevalence of intron definition in splicing in insects than in vertebrates (68,69). In plants, mean intron length does appear to correlate with total genome size, but mean intron length increases much more slowly with total genome size than in vertebrates. Similar correlations are observed between mean intron length and gene number, with a much more prominent difference between the slope of the correlation in plants and vertebrates (data not shown). The significance of this remains unclear.

Downloaded from https://academic.oup.com/nar/advance-article-abstract/doi/10.1093/nar/gkaa464/5850313 by guest on 05 June 2020

CONCLUDING REMARKS

We have created a database of intron annotation and homology information and used it to investigate several evolutionary hypotheses regarding the two classes of spliceosomal introns in eukaryotes. We have also created a web-based interface for querying this database to facilitate further investigations. The relationships between intron class, phase, terminal dinucleotides and -1 nucleotides at the 5′ splice site and the nonrandom distribution of U12-type introns annotated in the IAOD do not support many previous models that explain these patterns (6,30,50,51), but do support an extension of the class conversion model previously proposed (33).

DATA AVAILABILITY

The IAOD is publically accessible at introndb.lerner.ccf.org and all code used to create the database and run the website is available at the following GitHub repository https://github.com/Devlin-Moyer/IAOD. The standalone intronIC algorithm is available at https://github.com/glarue/intronIC.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

Michael Weiner provided invaluable technical support by hosting and managing the website. Rosemary Dietrich provided extensive background on splicing and general advice on various aspects of data interpretation. Daniel Blankenburg provided valuable feedback on various aspects of the web design and method of annotating orthologous introns. *Author contributions*: D.M. planned, designed and created the website and all scripts used to generate and process data except intronIC. G.L. designed and wrote intronIC. D.M. wrote the manuscript with input from all authors.

FUNDING

National Institutes of Health [R01GM104059, R01GM133989 to R.A.P.]; National Science Foundation [1616878, 1751372 to S.W.R.]. Funding for open access charge: NIH.
Conflict of interest statement. None declared.

REFERENCES

1. Turunen,J.J., Niemelä,E.H., Verma,B. and Frilander,M.J. (2013) The significant other: splicing by the minor spliceosome. *Wiley Interdiscip. Rev. RNA*, **4**, 61–76.
2. Chen,W. and Moore,M.J. (2014) The spliceosome: disorder and dynamics defined. *Curr. Opin. Struct. Biol.*, **24**, 141–149.
3. Russell,A.G., Charette,J.M., Spencer,D.F. and Gray,M.W. (2006) An early evolutionary origin for the minor spliceosome. *Nature*, **443**, 863–866.
4. Bartschat,S. and Samuelsson,T. (2010) U12 type introns were lost at multiple occasions during evolution. *BMC Genomics*, **11**, 106.
5. Hall,S.L. and Padgett,R.A. (1996) Requirement of U12 snRNA for in vivo splicing of a minor class of eukaryotic nuclear pre-mRNA introns. *Science*, **271**, 1716–1718.
6. Rogozin,I.B., Carmel,L., Csuros,M. and Koonin,E.V. (2012) Origin and evolution of spliceosomal introns. *Biol. Direct*, **7**, 11.
7. Dietrich,R.C., Incorvaia,R. and Padgett,R.A. (1997) Terminal intron dinucleotide sequences do not distinguish between U2- and U12-dependent introns. *Mol. Cell*, **1**, 151–160.
8. Zerbino,D.R., Achuthan,P., Akanni,W., Amode,M.R., Barrell,D., Bhai,J., Billis,K., Cummins,C., Gall,A., Girón,C.G. *et al.* (2018) Ensembl 2018. *Nucleic Acids Res.*, **46**, D754–D761.
9. Casper,J., Zweig,A.S., Villarreal,C., Tyner,C., Speir,M.L., Rosenbloom,K.R., Raney,B.J., Lee,C.M., Lee,B.T., Karolchik,D. *et al.* (2018) The UCSC Genome Browser database: 2018 update. *Nucleic Acids Res.*, **46**, D762–D769.
10. O'Leary,N.A., Wright,M.W., Brister,J.R., Ciufo,S., Haddad,D., McVeigh,R., Rajput,B., Robbertse,B., Smith-White,B., Ako-Adjei,D. *et al.* (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733–D745.
11. Lopez,P.J. and Séraphin,B. (2000) YIDB: the yeast intron dataBase. *Nucleic Acids Res.*, **28**, 85–86.
12. Sakharkar,M., Long,M., Tan,T.W. and de Souza,S.J. (2000) ExInt: an exon/intron database. *Nucleic Acids Res.*, **28**, 191–192.
13. Sakharkar,M.K., Kangueane,P., Woon,T.W., Tan,T.W., Kolatkar,P.R., Long,M. and de Souza,S.J. (2000) IE-Kb: intron exon knowledge base. *Bioinformatics*, **16**, 1151–1152.
14. Saxonov,S., Daizadeh,I., Fedorov,A. and Gilbert,W. (2000) EID: the exon-intron database-an exhaustive database of protein-coding intron-containing genes. *Nucleic Acids Res.*, **28**, 185–190.
15. Fedorov,A., Stombaugh,J., Harr,M.W., Yu,S., Nasalean,L. and Shepelev,V. (2005) Computer identification of snoRNA genes using a Mammalian Orthologous Intron Database. *Nucleic Acids Res.*, **33**, 4578–4583.
16. Bhasi,A., Philip,P., Manikandan,V. and Senapathy,P. (2009) ExDom: an integrated database for comparative analysis of the exon-intron structures of protein domains in eukaryotes. *Nucleic Acids Res.*, **37**, D703–D711.
17. Burset,M., Seledtsov,I.A. and Solovyev,V.V. (2001) SpliceDB: database of canonical and non-canonical mammalian splice sites. *Nucleic Acids Res.*, **29**, 255–259.
18. Alioto,T.S. (2007) U12DB: a database of orthologous U12-type spliceosomal introns. *Nucleic Acids Res.*, **35**, D110–D115.
19. Fedorov,A., Merican,A.F. and Gilbert,W. (2002) Large-scale comparison of intron positions among animal, plant, and fungal genes. *Proc. Natl. Acad. Sci. U.S.A.*, **99**, 16128–16133.
20. Fedorov,A., Roy,S., Fedorova,L. and Gilbert,W. (2003) Mystery of intron gain. *Genome Res.*, **13**, 2236–2241.
21. Chamary,J.-V. and Hurst,L.D. (2005) Biased codon usage near intron-exon junctions: selection on splicing enhancers, splice-site recognition or something else? *Trends Genet.*, **21**, 256–259.
22. Sakharkar,M.K., Tan,T.W. and de Souza,S.J. (2001) Generation of a database containing discordant intron positions in eukaryotic genes (MIDB). *Bioinformatics*, **17**, 671–675.
23. Shepelev,V. and Fedorov,A. (2006) Advances in the Exon-Intron Database (EID). *Brief. Bioinform.*, **7**, 178–185.
24. Szcześniak,M.W., Kabza,M., Pokrzywa,R., Gudyś,A. and Makałowska,I. (2013) ERISdb: a database of plant splice sites and splicing signals. *Plant Cell Physiol.*, **54**, e10.
25. Chorev,M., Guy,L. and Carmel,L. (2016) JuncDB: an exon-exon junction database. *Nucleic Acids Res.*, **44**, D101–D109.
26. Olthof,A.M., Hyatt,K.C. and Kanadia,R.N. (2019) Minor intron splicing revisited: identification of new minor intron-containing genes and tissue-dependent retention and alternative splicing of minor introns. *BMC Genomics*, **20**, 686.
27. Long,M., Rosenberg,C. and Gilbert,W. (1995) Intron phase correlations and the evolution of the intron/exon structure of genes. *Proc. Natl. Acad. Sci. U.S.A.*, **92**, 12495–12499.
28. Long,M., de Souza,S.J. and Gilbert,W. (1995) Evolution of the intron-exon structure of eukaryotic genes. *Curr. Opin. Genet. Dev.*, **5**, 774–778.
29. Gilbert,W. (1987) The exon theory of genes. *Cold Spring Harb. Symp. Quant. Biol.*, **52**, 901–905.
30. Dibb,N.J. and Newman,A.J. (1989) Evidence that introns arose at proto-splice sites. *EMBO J.*, **8**, 2015–2021.
31. Dibb,N.J. (1991) Proto-splice site model of intron origin. *J. Theor. Biol.*, **151**, 405–416.

Downloaded from https://academic.oup.com/nar/advance-article-abstract/doi/10.1093/nar/gkaa464/5850313 by guest on 05 June 2020

32. Sverdlov,A.V., Rogozin,I.B., Babenko,V.N. and Koonin,E.V. (2004) Reconstruction of ancestral protosplice sites. *Curr. Biol.*, **14**, 1505–1508.
33. Burge,C.B., Padgett,R.A. and Sharp,P.A. (1998) Evolutionary fates and origins of U12-type introns. *Mol. Cell*, **2**, 773–785.
34. Levine,A. and Durbin,R. (2001) A computational scan for U12-dependent introns in the human genome sequence. *Nucleic Acids Res.*, **29**, 4006–4013.
35. Sheth,N., Roca,X., Hastings,M.L., Roeder,T., Krainer,A.R. and Sachidanandam,R. (2006) Comprehensive splice-site analysis using comparative genomics. *Nucleic Acids Res.*, **34**, 3955–3967.
36. Lin,C.-F., Mount,S.M., Jarmołowski,A. and Makałowski,W. (2010) Evolutionary dynamics of U12-type spliceosomal introns. *BMC Evol. Biol.*, **10**, 47.
37. Cortes,C. and Vapnik,V. (1995) Support-vector networks. *Mach. Learn.*, **20**, 273–297.
38. Madan,V., Kanojia,D., Li,J., Okamoto,R., Sato-Otsubo,A., Kohlmann,A., Sanada,M., Grossmann,V., Sundaresan,J., Shiraishi,Y. *et al.* (2015) Aberrant splicing of U12-type introns is the hallmark of ZRSR2 mutant myelodysplastic syndrome. *Nat. Commun.*, **6**, 6042.
39. Niemelä,E.H. and Frilander,M.J. (2014) Regulation of gene expression through inefficient splicing of U12-type introns. *RNA Biol.*, **11**, 1325–1329.
40. Nojima,T., Rebelo,K., Gomes,T., Grosso,A.R., Proudfoot,N.J. and Carmo-Fonseca,M. (2018) RNA polymerase II phosphorylated on CTD Serine 5 interacts with the spliceosome during co-transcriptional splicing. *Mol. Cell*, **72**, 369–379.
41. Pineda,J.M.B. and Bradley,R.K. (2018) Most human introns are recognized via multiple and tissue-specific branchpoints. *Genes Dev.*, **32**, 577–591.
42. Cologne,A., Benoit-Pilven,C., Besson,A., Putoux,A., Campan-Fournier,A., Bober,M.B., De Die-Smulders,C.E.M., Paulussen,A.D.C., Pinson,L., Toutain,A. *et al.* (2019) New insights into minor splicing-a transcriptomic analysis of cells derived from TALS patients. *RNA*, **25**, 1130–1149.
43. Burke,J.E., Longhurst,A.D., Merkurjev,D., Sales-Lee,J., Rao,B., Moresco,J.J., Yates,J.R. 3rd, Li,J.J. and Madhani,H.D. (2018) Spliceosome profiling visualizes operations of a dynamic RNP at nucleotide resolution. *Cell*, **173**, 1014–1030.
44. Pedregosa,F. (2011) Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.
45. Durinck,S., Moreau,Y., Kasprzyk,A., Davis,S., De Moor,B., Brazma,A. and Huber,W. (2005) BioMart and bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, **21**, 3439–3440.
46. Durinck,S., Spellman,P.T., Birney,E. and Huber,W. (2009) Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.*, **4**, 1184–1191.
47. Buchfink,B., Xie,C. and Huson,D.H. (2015) Fast and sensitive protein alignment using DIAMOND. *Nat. Methods*, **12**, 59–60.
48. Larkin,M.A., Blackshields,G., Brown,N.P., Chenna,R., McGettigan,P.A., McWilliam,H., Valentin,F., Wallace,I.M., Wilm,A., Lopez,R. *et al.* (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947–2948.
49. Roy,S.W., Fedorov,A. and Gilbert,W. (2003) Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain. *Proc. Natl. Acad. Sci. U.S.A.*, **100**, 7158–7162.
50. Nguyen,H.D., Yoshihama,M. and Kenmochi,N. (2006) Phase distribution of spliceosomal introns: implications for intron origin. *BMC Evol. Biol.*, **6**, 69.
51. Long,M. and Rosenberg,C. (2000) Testing the 'proto-splice sites' model of intron origin: evidence from analysis of intron phase correlations. *Mol. Biol. Evol.*, **17**, 1789–1796.
52. Mount,S.M. (1982) A catalogue of splice junction sequences. *Nucleic Acids Res.*, **10**, 459–472.
53. Pomeranz Krummel,D.A., Oubridge,C., Leung,A.K.W., Li,J. and Nagai,K. (2009) Crystal structure of human spliceosomal U1 snRNP at 5.5 A resolution. *Nature*, **458**, 475–480.
54. Kondo,Y., Oubridge,C., van Roon,A.-M.M. and Nagai,K. (2015) Crystal structure of human U1 snRNP, a small nuclear ribonucleoprotein particle, reveals the mechanism of 5′ splice site recognition. *Elife*, **4**, e04986.
55. Turunen,J.J., Will,C.L., Grote,M., Lührmann,R. and Frilander,M.J. (2008) The U11-48K protein contacts the 5′ splice site of U12-type introns and the U11-59K protein. *Mol. Cell. Biol.*, **28**, 3548–3560.
56. Lynch,M. and Richardson,A.O. (2002) The evolution of spliceosomal introns. *Curr. Opin. Genet. Dev.*, **12**, 701–710.
57. Singh,J. and Padgett,R.A. (2009) Rates of in situ transcription and splicing in large human genes. *Nat. Struct. Mol. Biol.*, **16**, 1128–1133.
58. Basu,M.K., Rogozin,I.B. and Koonin,E.V. (2008) Primordial spliceosomal introns were probably U2-type. *Trends Genet*., **24**, 525–528.
59. Thanaraj,T.A. and Clark,F. (2001) Human GC-AG alternative intron isoforms with weak donor sites show enhanced consensus at acceptor exon positions. *Nucleic Acids Res.*, **29**, 2581–2593.
60. Farrer,T., Roller,A.B., Kent,W.J. and Zahler,A.M. (2002) Analysis of the role of Caenorhabditis elegans GC-AG introns in regulated splicing. *Nucleic Acids Res.*, **30**, 3360–3367.
61. Churbanov,A., Winters-Hilt,S., Koonin,E.V. and Rogozin,I.B. (2008) Accumulation of GC donor splice signals in mammals. *Biol. Direct*, **3**, 30.
62. Szafranski,K., Schindler,S., Taudien,S., Hiller,M., Huse,K., Jahn,N., Schreiber,S., Backofen,R. and Platzer,M. (2007) Violating the splicing rules: TG dinucleotides function as alternative 3′ splice sites in U2-dependent introns. *Genome Biol.*, **8**, R154.
63. Dietrich,R.C., Fuller,J.D. and Padgett,R.A. (2005) A mutational analysis of U12-dependent splice site dinucleotides. *RNA*, **11**, 1430–1440.
64. Vinogradov,A.E. (1999) Intron–Genome size relationship on a large evolutionary scale. *J. Mol. Evol.*, **49**, 376–384.
65. Deutsch,M. and Long,M. (1999) Intron-exon structures of eukaryotic model organisms. *Nucleic Acids Res.*, **27**, 3219–3228.
66. Patel,A.A. and Steitz,J.A. (2003) Splicing double: insights from the second spliceosome. *Nat. Rev. Mol. Cell Biol.*, **4**, 960–970.
67. Lynch,M. and Conery,J.S. (2003) The origins of genome complexity. *Science*, **302**, 1401–1404.
68. Fox-Walsh,K.L., Dou,Y., Lam,B.J., Hung,S.-P., Baldi,P.F. and Hertel,K.J. (2005) The architecture of pre-mRNAs affects mechanisms of splice-site pairing. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 16176–16181.
69. De Conti,L., Baralle,M. and Buratti,E. (2013) Exon and intron definition in pre-mRNA splicing. *Wiley Interdiscip. Rev. RNA*, **4**, 49–60.
70. Federhen,S. (2012) The NCBI Taxonomy database. *Nucleic Acids Res.*, **40**, D136–D143.
71. Letunic,I. and Bork,P. (2016) Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.*, **44**, W242–W255.

Downloaded from https://academic.oup.com/nar/advance-article-abstract/doi/10.1093/nar/gkaa464/5850313 by guest on 05 June 2020